Cross Parallax Attention Network for Stereo Image Super-Resolution

Canqiang Chen, Chunmei Qing, *Member, IEEE*, Xiangmin Xu, *Member, IEEE*, and Patrick Dickinson, *Member, IEEE*

***Abstract*—Stereo super-resolution (SR) aims to enhance the spatial resolution of one camera view using additional information from the other. Previous deep-learning-based stereo SR methods indeed improved the SR performance effectively by employing additional information, but they are unable to super-resolve stereo images where there are large disparities, or different types of epipolar lines. Moreover, in these methods, one model can only super-solve images of a particular view, and for one specific scale factor. This paper proposes a cross parallax attention stereo super-resolution network (CPASSRnet) which can perform stereo SR of multiple scale factors for both views, with a single model. To overcome the difficulties of large disparity and different types of epipolar lines, a cross parallax attention module (CPAM) is presented, which captures the global correspondence of additional information for each view, relative to the other. CPAM allows the two views to exchange additional information with each other according to the generated attention maps. Quantitative and qualitative results compared with the state of the arts illustrate the superiority of CPASSRnet. Ablation experiments demonstrate that the proposed components are effective and noise tests verify the robustness of CPASSRnet.**

***Index Terms*—Stereo super-resolution, single model for multiple scaling factors, attention mechanism, convolutional neural network.**

I. Introduction

STEREOSCOPIC images have many important applications, such as depth estimation, robot navigation, virtual reality (VR) or augmented reality (AR) displays and others. In recent years, many techniques have sought to solve the visual problems related to stereoscopic images, such as stereo deblurring [1], 3D object detection [2], object segmentation [3], stereo style transfer [4] and visual saliency prediction [5]. With the development of electronic devices (e.g., smartphones and 3D display devices) and hardware (e.g. CPUs and GPUs), the need for higher-quality and higher-resolution stereoscopic images has increased, giving rise to a demand for stereo super-resolution (SR) techniques. The maturity of Stereo SR technique can facilitate 3D content creators and produce better quality 3D resources. For example, the resolution of images that can be supported by different hardware conditions of 3D display devices is different. Through a simple stereo SR algorithm, the resolution of these images to be displayed can be dynamically scaled with high quality to adapt to each device.

The purpose of the single-image super-resolution (SISR)

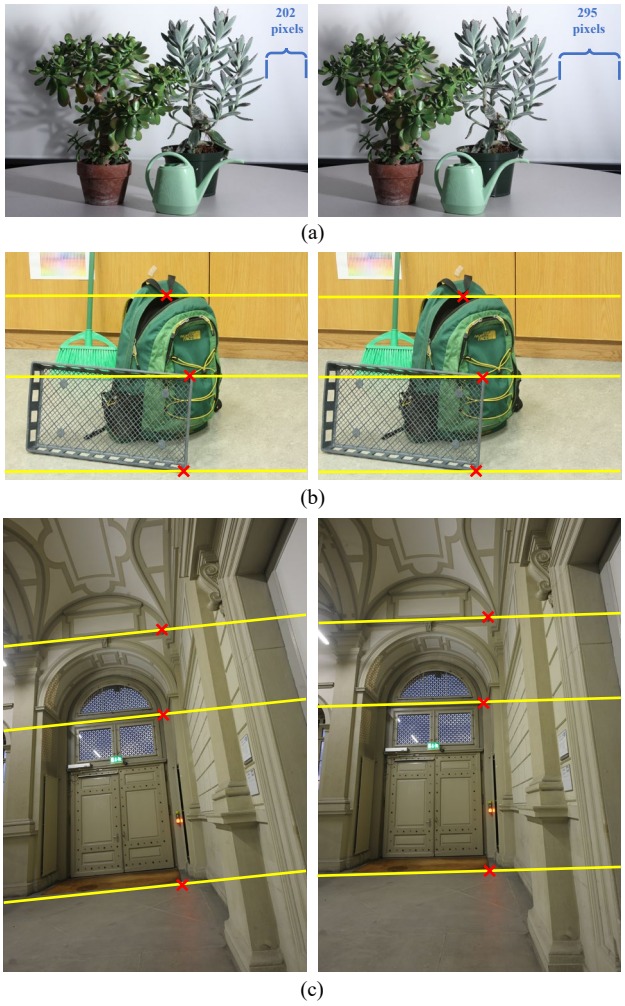

Fig. 1. (a) Stereo images with disparity of 93 pixels (larger than 64 pixels) from Middlebury [6] dataset. (b) Stereo images with horizontal epipolar line from Middlebury [6] dataset. (c) Stereo images with irregular epipolar line from ETH3D [7] dataset.

Manuscript received x x, 2020; This work is partially supported by the following grants: National Natural Science Foundation of China (61972163, U1801262, 61702192, U1636218, 61802131), and Science and Technology Project of Zhongshan (2019AG024).

C. Chen and C. Qing (corresponding author) are with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China (e-mail: canq.chen@gmail.com; qchm@scut.edu.cn).

X. Xu is with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China, and the Institute of Modern Industrial Technology of SCUT in Zhongshan, Zhongshan 528400, China (e-mail: xmxu@scut.edu.cn).

P. Dickinson is with the School of Computer Science, University of Lincoln, United Kingdom (e-mail: pdickinson@lincoln.ac.uk).

technology is to recover a high-resolution (HR) image from only a low-resolution (LR) image, while the purpose of multi-image super-resolution technology is to exploit the additional information provided by multiple LR views to recover the HR image corresponding to one of the views. These views were taken by cameras placed at different positions. Stereo image super-resolution is an example of multi-image super-resolution which uses two views, left and right. Due to the relative placement of the binocular cameras, parallax effects between the LR images causes a sub-pixel shift between them. Therefore, one view of LR images has additional information relative to the other. The SISR method is an ill-posed problem: it aims to recover a HR image from a single LR image without knowing the true value. The stereo SR is also an ill-posed problem, but it can use the additional information from another image, which makes the super-resolution result better than SISR. Therefore, the core of stereo SR is to find the correspondence of this additional information, and use it to enhance the stereo SR performance.

Recently, a convolutional neural network, term StereoSR, was proposed by Jeon *et al.* [8] to perform super-resolution task for stereo images. This network concatenated, as input, a left image and shifted right images, with different intervals, so the network can learn the correspondence from the concatenated image batch. However, they set a prior that the parallax shift was limited to 64 pixels, making it unsuitable for images with larger disparity. Aiming to overcome this limitation, Wang *et al.* [9] leveraged the attention mechanism to find the correspondence along the horizontal epipolar line, without disparity limit. More recently, Ying *et al.* [10] proposed a stereo attention module (SAM) and inserted it into the SISR models to perform stereo SR task. However, its attention mechanism is the same as that of [9]. These two algorithms can only super-resolve the stereo images captured by the binocular cameras in standard form (epipolar line is horizontal). However, stereo images can have irregular epipolar lines, and have large disparity, as shown in Fig. 1. These methods improve the stereo SR performance in different meaningful ways, but they are unable to deal with all cases.

In addition, they can only perform SR of one scale factor, in one model, which is not in line with the current pursuit: single model, multiple scale factors. Many applications in real scenarios require solving super-resolution tasks of different scale factors, such as scaling images to fit displays of different resolutions. In order to achieve this kind of multitasking application, single-scale models require repeated training, which is inefficient and time-consuming, and lacks flexibility for future use. Therefore, it is very advantageous in terms of resources and convenience to train a multi-scale model that can handle multiple SR tasks of different scale factors.

In order to perform stereo SR of multiple scale factors for both views with a single model, this paper proposes a cross parallax attention stereo SR network (CPASSRnet) for stereo SR task. The overall network is an encoder-decoder architecture whose benefits are as follows. First, it has large enough receptive field to find a better global stereo correspondence. Second, it can be trained by interpolated LR samples with different upscaling factors, so that a single network is able to perform the SR task for multiple upscaling factors. To facilitate the interaction of additional information between the left and right views, we propose a cross parallax attention module (CPAM) that globally captures the stereo correspondence of respective additional information, without a maximum disparity limit or epipolar line limit. The CPAM enables the network to process the left and right views simultaneously in one feed. The superiority of CPAM and CPASSRnet was verified by extensive experiments.

The contribution of this paper is fourfold:

1) We propose a novel network named as CPASSRnet for the stereo SR problem, which can super-resolve the left and right views in one feed, and can perform SR of multiple upscaling factors.

2) A cross parallax attention module is proposed to capture the global correspondence of additional information between the stereo images to reconstruct more details during super-resolution. It can successfully process stereo pairs with large disparity variations, and different types of epipolar lines.

3) A modified version of perceptual loss is proposed to better suit the stereo SR problem, and to obtain more perceptually realistic and accurate SR images. A multi-scale training strategy is used for boosting the SR performance.

4) Extensive experiments illustrate that the proposed CPASSRnet network can restore high-resolution images with higher visual quality, and obtain higher PSNR/SSIM scores in quantitative evaluation, compared to representative single image SR and state-of-the-art stereo SR baselines. The noise test also shows that our method has stronger anti-noise ability than state-of-the-art stereo SR methods.

II. Related Work

A. Single Image Super-Resolution

Before the deep learning era, many classic methods have been proposed to solve the single image super-resolution problem, such as edge-based methods [11], [12], patch-based methods [13]–[15], self-example-learning-based [16], [17], statistics-based methods [18], [19] and sparse-representation-based methods [20], [21]. Dong *et al.* [22] proposed a SRCNN network, and were the first to apply a Convolution Neural Network (CNN) to solve the SR task: they achieved better performance than traditional methods. To improve the performance of SISR, many CNN-based methods were then further proposed. Ledig *et al.* [23] leveraged Generative Adversarial Network (GAN) [24] to generate photo-realistic SR images. A new sub-pixel upsampling method was proposed by Shi *et al.* [25] to reduce the computational complexity, which is used by many subsequent methods. Methods like VDSR [26], SRDenseNet [27], EDSR [28], RDN [29], CARN [30] and MSRN [31] applied residual learning [32] or dense skip connections to improve SR performance. Recently, Zhang *et al.* [33] developed a residual in residual network, term RCAN, for accurate SISR task, which was also powered by a channel attention mechanism. RFANet [34] extended RCAN's residual network by aggregating the residual features and achieved better results. Tian *et al.* [35] proposed to collect more complementary contextual information to improve the performance

of SISR, in a way of coarse-to-fine. Guo *et al.* [36] designed a dual-view attention networks: local aware attention for LR feature space and global aware attention for HR feature space, making effective SISR results. Combining non-local attention mechanism [37], Mei *et al.* [38] modeled cross-scale feature correlation to find similar patches in the same LR image to improve the SR quality of target patch. Most recently, a holistic attention network which achieved a favorable performance was proposed by Niu *et al.* [39]. They jointly applied layer attention and channel-spatial attention, so that the network was able to capture comprehensive interdependencies, including positions, channels and layers. The SISR methods recover HR image from only a single LR image without knowing the ground truth, which is ill-posed. In order to get a better solution, more information needs to be integrated. The stereo SR methods are from this aspect to improve SR performance.

B. Stereo Image Super-Resolution

Conventional stereo SR methods [40], [41] reconstructed the HR images using the parallax information estimated by traditional depth estimation algorithms, and registered the pixels in a stereo pair based on the that parallax. Such methods were limited by the inaccuracy of the estimation algorithms, and thus did not perform very well. Jeon *et al.* [8] designed a StereoSR network that contains two subnetworks to enhance the spatial resolution of the left image. They shifted the right image horizontally, by different number of pixels to obtain multiple shifted right images which provide correspondence cues for the left image. But their method has application limitations, because once the network is trained, the maximum parallax of the stereo image that can be processed is fixed (64 in the original paper). To this end, a system called PASSRnet was exploited by Wang *et al.* [9] for stereo image pairs, with all possible disparities. They extended self-attention and proposed a parallax attention module (PAM) for stereo SR. Then a stereo attention module (SAM) was proposed by Ying *et al.* [10] to extend SISR to stereo SR. But its mechanism is the same as PAM, which just extends PAM to handle the left and right views at the same time. Song *et al.* [42] aimed to maintain consistent texture details between the super-resolved left and right views. They incorporated self-attention into PAM and developed a self and parallax attention mechanism (SPAM) to aggregate own image information and stereo image counterpart information. Recently, Yan *et al.* [43] solved stereo SR through another way. They designed a multi-task network that first predicted the disparity map based on the input stereo pair. Then the predicted disparity map were taken as a prior to adatively facilitate the capture of cross-view information, so as to better resotre images. Although PAM, SAM and SPAM can find the correspondence along the horizontal epipolar line without disparity limit, they cannot super-resolve the stereo images captured by a non-standard form of binocular system (e.g. dual cameras arranged vertically on a smartphone). Furthermore, these models are single-scale, and one model needs to be trained for each specific scale factor. Our method uses a multi-scale training strategy, so that one model can handle multiple scale factors.

C. Attention Mechanism

Many scholars have applied self-attention to multiple fields, including natural language processing [44], [45], image synthesis [46] and image recognition [37]. In computer vision tasks, it learns the dependency between any two pixels within the spatial dimension of an input image to find global dependencies. It also was extended to generate an attention map (or mask with weights) for scaling the deep features. Therefore, the attention map contained the dependencies between two spatial pixels, or two channels, or two feature maps. SENet [47] and SKNet [48] utilized the attention mechanism to model the channel-wise importance, to rescale the channel features for image classification tasks. In addition to channel attention, CBAM [49] jointly employed spatial attention as well as channel attention to enhance the feature representation ability of CNN. Since previous attention modules often only utilize first-order statistics of features, the statistics of features higher than first order maight be ignored. Hence, a second-order attention network [50] was proposed to enhance the network's discriminative ability, and achieved impressive performance in handling SISR. Recently, Qiao *et al.* [51] presented hierarchical network for image matting: They leveraged spatial-wise attention to filter appearance cues, and channel-wise attention to distill pyramidal features, which enables the network to perceive better alpha mattes. In object detection and semantic segmentation tasks, Yin *et al.* [52] found through in-depth research that the attention calculation process of the non-local block [37] should be divided into two parts, otherwise it would hinder the overall network performance. Thereby they designed a disentangled non-local block which contains two parts: a whitened pairwise term that calculates the relationship between every two pixels, and unary term that represents every pixel's saliency. The effectiveness of this disentangled block was verified. In addition to image tasks, Li *et al.* [53] and Yan *et al.* [54] proposed to use spatio-temporal attetion to complete video action recognition and video annotation tasks. In summary, by modeling the correlation and interdependence of deep features, attention mechanism can improve the performance of many low-level vision tasks.

III. The Cross Parallax Attention Stereo Super-Resolution Network

In this section, we will describe in detail our proposed network. We first briefly introduce the overall network architecture, and then illustrate the details of the network, including the components in the network and the downsampling /upsampling method, etc. At last, we present the proposed core cross parallax attention module and loss functions in detail.

A. Network Architecture

In the following, we denote low resolution (LR) stereo left and right images as $\mathbf{I}^l_{LR} \in \mathbb{R}^{sH\times sW\times 3}$, $\mathbf{I}^r_{LR} \in \mathbb{R}^{sH\times sW\times 3}$, respectively. The notation s is used to denote the downscaling factor (e.g., 1/2, 1/3 and 1/4), and H and W represent the desired image height and width, respectively. Then the LR stereo images are upscaled to the desired resolution, denoted as $\mathbf{I}^l_{ILR} \in \mathbb{R}^{H\times W\times 3}$ and $\mathbf{I}^r_{ILR} \in \mathbb{R}^{H\times W\times 3}$, by bicubic

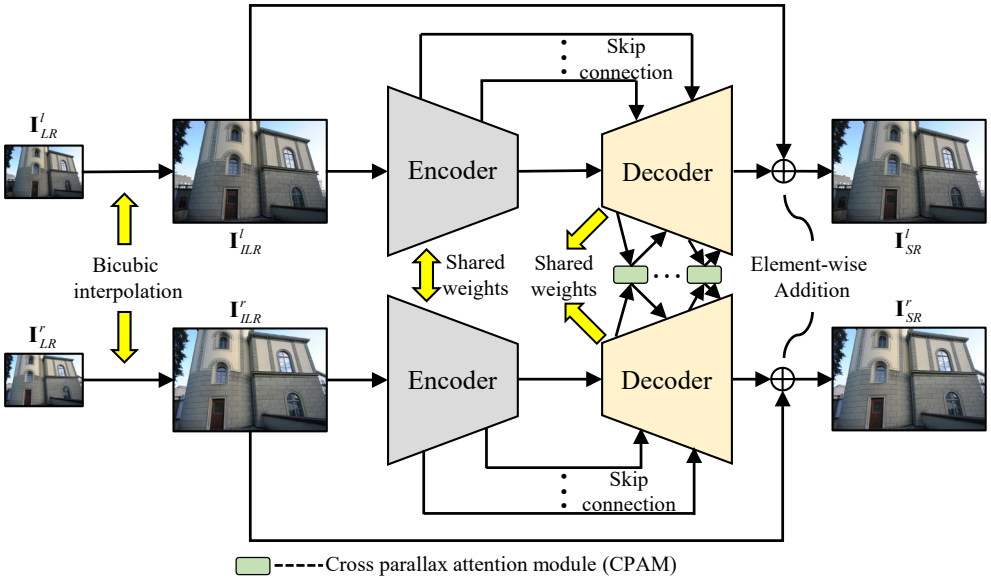

Fig. 2. Network architecture of CPASSRnet. The LR stereo image pair ($\mathbf{I}^l_{LR}$, $\mathbf{I}^r_{LR}$) is first upsampled bicubically to the desired spatial size to obtain the interpolated image pair ($\mathbf{I}^l_{ILR}$, $\mathbf{I}^r_{ILR}$). Then $\mathbf{I}^l_{ILR}$ and $\mathbf{I}^r_{ILR}$ are fed into two autoencoder networks that share parameters, and there are multiple skip connections between the encoder and the decoder for feature reuse. Multiple cross parallax attention modules (CPAM) are set between the decoders of the two branches to capture the stereo correspondence between the left and right views and thereby enhance the features. Finally, the interpolated image pair and the decoder output are directly added to obtain the final output.

interpolation [55]. To obtain a multi-scale model, we adopt a multi-scale training strategy which inputs images interpolated at different scale factors into the network for training.

The overall architecture of CPASSRnet is shown in Fig. 2, which is designed as an encoder-decoder end-to-end network. The network takes a interpolated low-resolution (ILR) stereo image pair $(\mathbf{I}^l_{ILR}, \mathbf{I}^r_{ILR})$ as input, and outputs the super-resolved high-resolution stereo image pair $(\mathbf{I}^l_{SR}, \mathbf{I}^r_{SR})$, $\mathbf{I}^l_{SR} \in \mathbb{R}^{H\times W\times 3}$, $\mathbf{I}^r_{SR} \in \mathbb{R}^{H\times W\times 3}$. The overall network consists of two branches, one for left images and the other for right images, and both of the branches share parameters. First, the Encoder processes the ILR images to extract multi-scale features, and then the Decoder reuses these multi-scale features to recover higher spatial resolution Decoder output. Finally, the input ILR images are directly added to the Decoder output, generating the super-resolved output images. Multi-stage additional information interaction between the left view and right view is realized between the two Decoders, by setting up multiple our proposed cross parallax attention module (CPAM) which is described in Sec. III-D, and thus the details of the recovered SR images are enhanced.

This encoder-decoder architecture is employed as the backbone network because when the spatial resolution of the feature maps decreases, the receptive field of the model increases. This characteristic is highly suitable for stereo SR problems. Since there is a certain parallax between the left and the right views, the model needs a large enough receptive field to find the global stereo correspondence of one view on another. In addition, such architecture takes interpolated images (upscale to desire spatial resolution) as input, which enables a single network to perform the stereo SR task for multiple scale factors.

B. Encoder/Decoder Design

The Encoder encodes the input stereo image $\mathbf{I}^l_{ILR}$ or $\mathbf{I}^r_{ILR}$ and generates multi-scale feature maps with scaling step of 2 as shown in Fig. 3. Since the spatial resolution of the feature map decreases by a factor of 2, the channel number will correspondingly increase by a factor of 2. These encoded left and right multi-scale feature maps are respectively represented as $\mathbf{F}^{el}_i \in \mathbb{R}^{H_i\times W_i\times C_i}$ and $\mathbf{F}^{er}_i \in \mathbb{R}^{H_i\times W_i\times C_i}$, where $i \in \{1, 2, 3, \cdots, K\}$, and K represents the total number of downsampling layers (K is set to 4 by the sensitivity analysis in Sec. IV-C). H_i, W_i and H_i are used to denote the height, width and channel of the i-th downsampling layer's output feature map, respectively. The spatial size relationship between $\mathbf{F}^{el}_i$/$\mathbf{F}^{er}_i$ and $\mathbf{F}^{el}_{i+1}$/$\mathbf{F}^{er}_{i+1}$ is given by Eq. 1:

$$H_i = 2 \times H_{i+1}, W_i = 2 \times W_{i+1}, C_i = \frac{1}{2} \times C_{i+1}. \quad (1)$$

The Decoder is symmetrical with the Encoder. It starts with the final encoded feature map and gradually recovers spatial resolution using an upsampling module. Similarly, the left and right feature maps in the Decoder are respectively represented by $\mathbf{F}^{dl}_j \in \mathbb{R}^{H_j\times W_j\times C_j}$ and $\mathbf{F}^{dr}_j \in \mathbb{R}^{H_j\times W_j\times C_j}$, where $j \in \{1, 2, 3, \cdots, K\}$, K represents the total number of upsampling modules (the same as the downsampling layers), and H_j, W_j and H_j are the height, width and channel of the j-th upsampling module's output feature map, respectively. The spatial size relationship between $\mathbf{F}^{dl}_j$/$\mathbf{F}^{dr}_j$ and $\mathbf{F}^{dl}_{j+1}$/$\mathbf{F}^{dr}_{j+1}$ is given by Eq. 2:

$$H_j = \frac{1}{2} \times H_{j+1}, W_j = \frac{1}{2} \times W_{j+1}, C_j = 2 \times C_{j+1}. \quad (2)$$

To encourage multi-scale features reuse and aggregate more useful information, each earlier Encoder feature map is fed

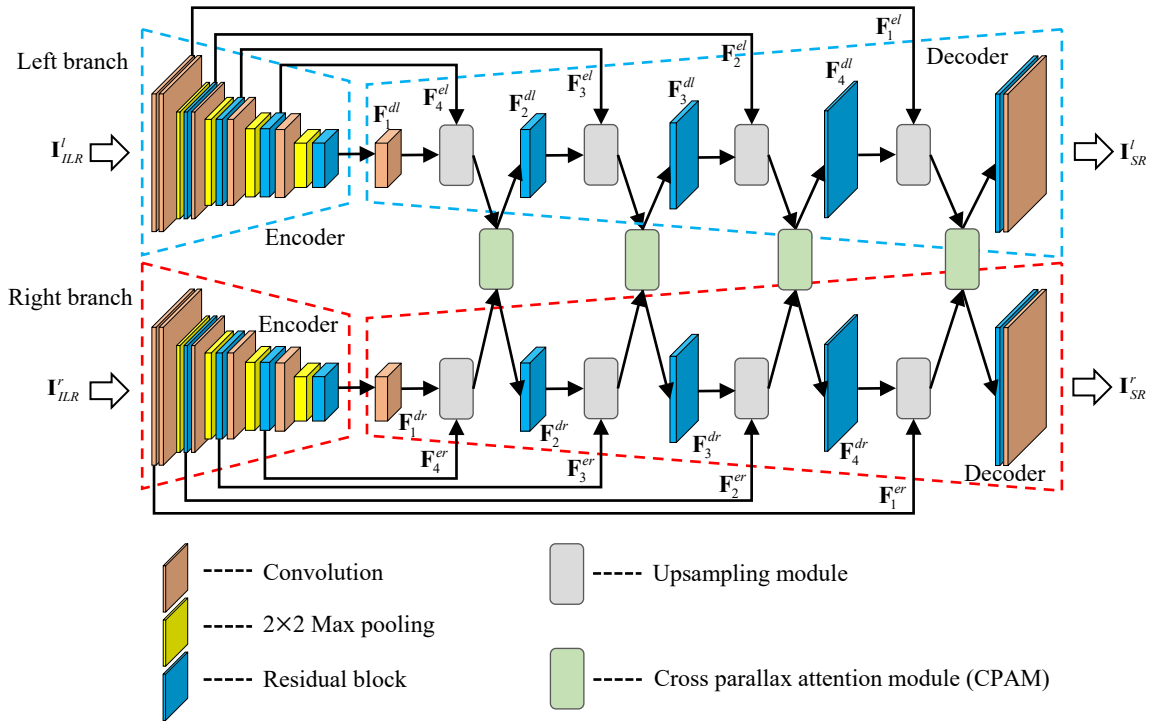

Fig. 3. Details of the encoder-decoder structure. In this auto-encoder network, the encoder utilizes max pooling layers to step-downsample the input image, and then the decoder exploits the upsampling module to step-upsample the output of the encoder. The proposed CPAM detects and exchanges additional information between the feature maps of the left and right views. Each convolutional layer in the entire network is activated by rectified linear unit (ReLu).

into the upsampling module to be merged with the corresponding Decoder feature map. Each upsampling module is followed by a cross parallax attention module. Moreover, the proposed CPASSRnet utilizes residual block [22] in both the Encoder and Decoder, to enhance the representative ability of the network.

C. Downsampling /Upsampling Method

We utilized max pooling as our downsampling method, with pooling size of 2×2 and stride of 2, for the following reasons. First, max pooling does not include addition and multiplication operations, which can reduce the computational cost compared with stride-2 convolution. Secondly, max pooling can suppress noise, making the network more robust. The anti-noise effect can be demonstrated by the noise test in Sec. IV-G.

As shown in Fig. 3, Decoder exploits the upsampling module to up-sample the final encoded feature map, step-by-step, until the desired spatial resolution (the same as the input). The upsampling module is the intersection of the Encoder feature and the corresponding Decoder feature. Multiple skip connections from Encoder to Decoder can alleviate the problem of disappearing gradients, accelerate network convergence, fully reuse previously extracted features and encourage the flow of information between different layers [32].

Each upsampling module takes two feature maps as input, one from the corresponding Encoder layer output and one from the previous Decoder layer output as shown in Fig. 4 (illustrating the j-th upsampling module as an example). The feature map $\mathbf{F}_j^{dl}/\mathbf{F}_j^{dr}$ from the previous Decoder layer is first upscaled by $2\times$ bilinear. Then it undergos a 3×3 convolution

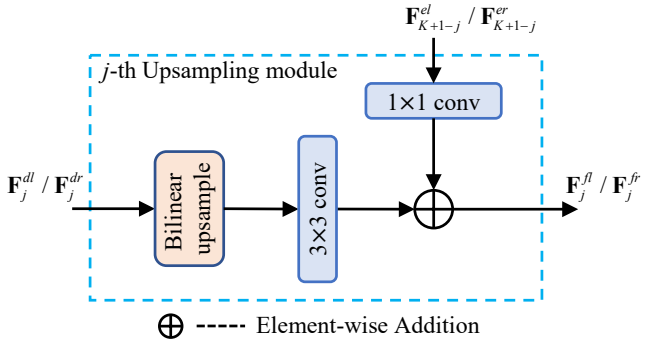

Fig. 4. Illustration of the upsampling module. The upsampling module first performs bilinear upsampling on the input decoder feature maps $\mathbf{F}_j^{dl}/\mathbf{F}_j^{dr}$, and fuses the upsampled feature maps with the encoder feature maps $\mathbf{F}^{el}_{K+1-j}/\mathbf{F}^{er}_{K+1-j}$ by element-wise addition as output.

operation, reducing the aliasing effect of interpolation as well as reducing the feature channels by half. Meanwhile, the feature map $\mathbf{F}^{el}_{K+1-j}/\mathbf{F}^{er}_{K+1-j}$ from the Encoder undergoes a 1×1 smooth transition convolution, without changing channel number. At this point, both feature maps have the same spatial resolution and channel number. These two processed feature maps are finally merged as module output by element-wise addition. The whole process can be formulated by Eq. 3:

$$\begin{cases}\mathbf{F}_j^{fl} = \mathbf{W}_j^1 * \mathbf{F}^{el}_{K+1-j} + \mathbf{W}_j^3 * BI^2(\mathbf{F}_j^{dl}) \\ \mathbf{F}_j^{fr} = \mathbf{W}_j^1 * \mathbf{F}^{er}_{K+1-j} + \mathbf{W}_j^3 * BI^2(\mathbf{F}_j^{dr}),\end{cases} \tag{3}$$

where $\mathbf{W}_j^1$ and $\mathbf{W}_j^3$ are the 1×1 convolutional kernel, and 3×3 convolutional kernel, in j-th upsampling module respectively,

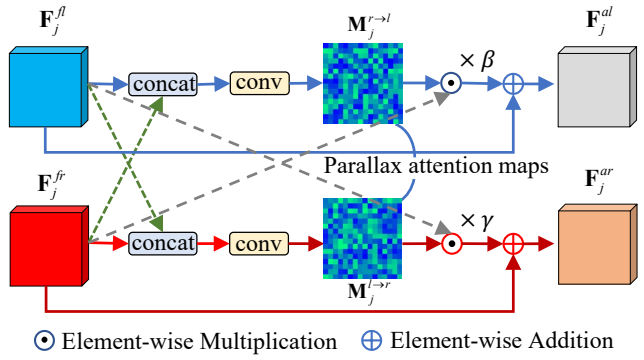

Fig. 5. The proposed Cross Parallax Attention Module (CPAM). CPAM first concatenates the input fused left feature map $\mathbf{F}_j^{fl}$ and fused right feature map $\mathbf{F}_j^{fr}$, and then convolves the concatenated feature maps to detect the stereo correspondence, yielding the attention maps $\mathbf{M}_j^{r\to l}$ and $\mathbf{M}_j^{l\to r}$ that characterize the stereo correspondence. Finally, according to these attention maps, the extra information on one view is integrated into another view.

$\mathbf{F}_{K+1-j}^{el}$/$\mathbf{F}_{K+1-j}^{er}$ is the $(K+1-j)$-th Encoder feature map, $\mathbf{F}_j^{dl}$/$\mathbf{F}_j^{dr}$ is the corresponding j-th Decoder feature map, $\mathbf{F}_j^{fl}$ and $\mathbf{F}_j^{fr}$ are fused feature maps output by the j-th upsampling module, $*$ denotes the convolution operation, and $BI^2(\cdot)$ represents the $2\times$ bilinear operation.

D. Cross Parallax Attention Module

Attention mechanisms have made significant improvements in many tasks [37], [44], [46] by learning the interdependence between the deep features, and generating masks (or attention maps) to re-weight those features. This mechanism is highly suitable for stereo SR problem, and enhances the spatial resolution of one view with additional information from another, helping to recover more detail. In our work we use a cross parallax attention module (CPAM) to jointly capture global interdependencies between stereo images in feature space, for multiple stages.

As shown in Fig. 5, the j-th CPAM takes two fused feature maps $\mathbf{F}_j^{fl} \in \mathbb{R}^{H_j\times W_j\times C_j}$ and $\mathbf{F}_j^{fr} \in \mathbb{R}^{H_j\times W_j\times C_j}$ as input, where H_j, W_j are the feature map height and width, and C_j is the feature map channel. These two feature maps are generated from the previous j-th upsampling module, and we take the left branch as an example of how the CPAM processes the fused features $\mathbf{F}_j^{fl}$ and $\mathbf{F}_j^{fr}$. The fused left feature $\mathbf{F}_j^{fl}$ and the fused right feature $\mathbf{F}_j^{fr}$ are concatenated first, and then the concatenated feature map is fed into a convolution (activated by ReLu [56]) with a larger size of 7×7. The 7×7 convolution automatically generalizes the extra information on the right feature from the concatenated feature map by virtue of its larger receptive field, resulting in a parallax attention map $\mathbf{M}_j^{r\to l} \in \mathbb{R}^{H_j\times W_j\times 1}$ which contains the stereo correspondence of the left feature with the right feature. Consequently, the additional information from the right feature to the left feature is represented by the element-wise product of $\mathbf{M}_j^{r\to l}$ and $\mathbf{F}_j^{fr}$. The additional information is finally incorporated into $\mathbf{F}_j^{fl}$ by element-wise addition of the previous product, and the left feature map $\mathbf{F}_j^{fl}$, producing the output attended left feature map $\mathbf{F}_j^{al} \in \mathbb{R}^{H_j\times W_j\times C_j}$. Before the final element-wise addition, a trainable parameter β is defined to adaptively scale the previous product, allowing better information incorporation. The attention process of the left feature can be represented as:

$$\begin{cases} \mathbf{M}_j^{r\to l} = \mathbf{W}_j^{l7} * cat(\mathbf{F}_j^{fl}, \mathbf{F}_j^{fr}) \\ \mathbf{F}_j^{al} = (\mathbf{M}_j^{r\to l} \odot \mathbf{F}_j^{fr}) \times \beta + \mathbf{F}_j^{fl}, \end{cases} \tag{4}$$

where $\mathbf{W}_j^{l7}$ is the 7×7 convolution kernel in the j-th CPAM for the left feature, $\odot$ represents the element-wise multiplication and $cat(\cdot)$ denotes the concatenation operation.

By the same inference, the attention process representation of the right view can be obtained by:

$$\begin{cases} \mathbf{M}_j^{l\to r} = \mathbf{W}_j^{r7} * cat(\mathbf{F}_j^{fr}, \mathbf{F}_j^{fl}) \\ \mathbf{F}_j^{ar} = (\mathbf{M}_j^{l\to r} \odot \mathbf{F}_j^{fl}) \times \gamma + \mathbf{F}_j^{fr}, \end{cases} \tag{5}$$

where $\mathbf{M}_j^{l\to r}$ is a parallax attention map that contains the left view's correspondence and parallax information with respect to the right view, $\mathbf{W}_j^{r7}$ is the 7×7 convolution kernel in the j-th CPAM for the right feature, γ is a trainable scaling factor and $\mathbf{F}_j^{ar}$ is the output attended right feature map, respectively.

E. Loss Function

In this section, two loss functions are defined to facilitate the network training and improve the stereo SR performance. These are the SR reconstruction loss, and L1-perceptual loss.

SR Reconstruction Loss. Instead of using mean square error (MSE) like previous stereo SR methods [8], [9], mean absolute error (MAE) is leveraged to reconstruct HR images, avoiding overly smooth textures:

$$\mathcal{L}_{rec} = \|\mathbf{I}_{SR}^{l} - \mathbf{I}_{HR}^{l}\|_1 + \|\mathbf{I}_{SR}^{r} - \mathbf{I}_{HR}^{r}\|_1, \tag{6}$$

where $\mathbf{I}_{HR}^{l}$ and $\mathbf{I}_{HR}^{r}$ represent the HR ground truth of the left view and right view, respectively.

L1-perceptual Loss. The VGG19 network [57] is proposed for image classification, and many SISR algorithms like [23] and [58] employed it to calculate the perceptual loss which evaluates the perceptual similarity between the super-resolved image and the ground truth. We define a modified perceptual loss named L1-perceptual loss which takes the shallow features of the pre-trained VGG19 network to calculate the perceptual loss, since calculating the loss using deeper features may give rise to visible artifacts and distortions. The shallow features calculated by L1-perceptual loss are generated by $relu1\_1$, $pool1$ $conv2\_2$ in the VGG19 network, not just the features generated by the ReLu layers. Such an approach makes the details of the super-resolved images more complete, because it exploits the features derived from different types of layers for calculation. The formulation of the L1-perceptual loss is:

$$\begin{aligned} \mathcal{L}_{per} = \frac{1}{H_t W_t C_t} \sum_{k=1}^{C_t} \sum_{j=1}^{W_t} \sum_{i=1}^{H_t} \Big(& \left\| \phi_{i,j,k}^{t}(\mathbf{I}_{SR}^{l}) - \phi_{i,j,k}^{t}(\mathbf{I}_{HR}^{l}) \right\|_1 \\ & + \left\| \phi_{i,j,k}^{t}(\mathbf{I}_{SR}^{r}) - \phi_{i,j,k}^{t}(\mathbf{I}_{HR}^{r}) \right\|_1 \Big), \end{aligned} \tag{7}$$

where $\phi_{i,j,k}^{t}$ indicates the (i, j, k) location of the feature map generated by layer t (can be $relu1\_1$, $pool1$ or $conv2\_2$), and

H_t, W_t and C_t are height, width and channel of layer t's feature map.

Total Loss. Equation 8 is the total loss function used to train the network:

$$\mathcal{L} = \lambda \mathcal{L}_{rec} + \eta \mathcal{L}_{per}, \tag{8}$$

where λ and η are hyper-parameters that balance the SR reconstruction loss and the L1-perceptual loss.

F. Differences Between CPASSRnet and PASSRnet

Firstly, PASSRnet is a post-upsampling method, which extracts features and reconstructs high-frequency components in LR space, and then reconstructs to HR space through pixel-shuffle [25] upsampling to obtain super-resolved images. However, CPASSRnet is a pre-upsampling method, which first upsamples the images in LR space to ILR space using traditional interpolation algorithms, and then inputs them to the network to obtain super-resolved images. Therefore, PASSRnet mainly manipulates images in LR space, while CPASSRnet mainly manipulates images in ILR space.

Secondly, the parallax-attention module (PAM) [9], which is used in PASSRnet, utilizes epipolar constraints to search for additional information along the horizontal epipolar lines, and this horizontal prior is only suitable for images captured by a stereo camera in standard form. Furthermore, PAM can only enhance one view (left or right) once it is trained, so it requires two SR model, one for each view. In contrast to PAM, our proposed CPAM can automatically capture the mutual stereo correspondence on the whole stereo image, and is not limited to the horizontal direction. Therefore, CPAM can enhance both views simultaneously using the captured mutual stereo correspondence.

In addition, PASSRnet deploys only one PAM in LR space before the post-upsampling, which may lead to insufficient information enhancement. In contrast, CPASSRnet sets multiple CPAMs in the Decoder, so the network can refine the images for multiple stages during the progressive recovery process, which makes the information interaction more adequate.

IV. Experiments

In this Section, the details of the training set as well as the test set and the training settings are fisrt presented. Next, we conduct sensitivity analysis to determine the suitable number of downsampling layers and ablate the cross parallax attention module as well as the L1-perceptual loss to examine their effectiveness. Then, on images with horizontal epipolar line, we evaluate the performance of CPASSRnet and with several baselines through multiple aspects (including quantitative scores, visual quality, computational time and the ability to find stereo correspondence). In addition, we explore the ability of CPASSRnet and the state of the arts in stereo SR to handle the stereo images with irregular epipolar lines. At last, we conduct noise test to demonstrate the noise immunity of CPASSRnet.

A. Image dataset

Following the experimental setting of PASSRnet [9] and StereoSR [8], the Middlebury [6], [59], [60] dataset is used to train the proposed CPASSRnet. Since the spatial resolution of images from Middlebury 2014 dataset [60] is much bigger than that from Middlebury 2003 dataset [59] and Middlebury 2006 dataset [6], the Middlebury 2014 dataset was down-sampled by $\frac{1}{2}\times$ bicubic interpolation to match the spatial resolution of the other two datasets. To test the performance on stereo images with horizontal epipolar lines, we seleted 5 images from Middlebury ('Cloth2', 'Motorcycle', 'Piano', 'Pipes' and 'Sword2') and randomly picked 100 images from each of the KITTI 2012 [61] and the KITTI 2015 [62] datasets to constructed an image test set. To test the performance on stereo images with irregular epipolar line, we randomly picked 5 stereo pairs from ETH3D [7] as a test set.

B. Implementation Details

We cropped 27160 training samples from the image training set. Images patches were 96×288 pixels in size, and non-overlapping. These samples were down-sampled by bicubic interpolation to generate LR patches which were then upscaled again as ILR patches by bicubic interpolation to be fed into the model. Since this study aims to perform stereo SR task for multiple scale factors using a single model, we adopt a multi-scale training strategy of batching the ILR patches at different scale factors ($\times 2$, $\times 3$ and $\times 4$) together to train the network. The data augmentation method was to randomly and horizontally flip the patches. Peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) on RGB space were used to quantitatively evaluate the super-resolved results.

We used the following default training settings. The Adam optimizer [63] with a batch size of 6 (contains patches for different scale factors), a learning rate of 0.0001, and exponential decay rates $(\beta_1, \beta_2) = (0.9, 0.999)$ was set to train the model. Hyper-parameters are set by sensitivity analysis as follows: $\lambda = 10$, $\eta = 0.1$, and the number of downsampling layers is set by sensitivity analysis to 4. The CPASSRnet models in subsequent experiments were implemented using Tensorflow and were trained for 1.5×10^5 iterations with learning rate linearly decay after 7.5×10^4 iterations; the model training was deployed on a single NVIDIA GTX 1080Ti. The source code released at: https://github.com/canqChen/CPASSRnet.

C. Sensitivity Analysis

In this part, we will conduct experiments to determine the hyper-parameters, including the number of downsampling layers and loss weights λ, η.

1) Determination of the Number of Downsampling Layers: In CPASSRnet, the number of downsampling layers N determines the number of different feature scales. Since the number of downsampling layers is fixed once the network is designed, too many downsampling layers will reduce the deep feature map's spatial resolution, giving rise to weaker feature representation. On the other hand, too few downsampling layers will limit the diversity of features from different scales, leading to an inability to enrich the details of the SR results.

TABLE I
MEAN PSNR (DB)/SSIM VALUES ACHIEVED ON MIDDLEBURY TEST SET (5 IMAGES) BY CPASSRNET FOR DIFFERENT SCALE FACTORS. NOTE THAT THE VALUE MARKED IN **RED** IS THE HIGHEST, AND THE VALUE IN **BLUE** IS THE SECOND HIGHEST. THE SAME SCHEME IS USED FOR SUBSEQUENT TABLES AND FIGURES.

N	3		4		5	
Scale	Left	Right	Left	Right	Left	Right
×2	**39.76**/**0.976**	**39.42**/**0.976**	**39.76**/**0.977**	**39.41**/**0.976**	**39.52**/**0.975**	39.17/**0.975**
×3	**35.48**/**0.938**	**35.38**/**0.939**	**35.63**/**0.938**	**35.49**/**0.939**	35.44/**0.936**	35.31/**0.937**
×4	33.16/**0.898**	33.14/**0.900**	**33.37**/**0.900**	**33.30**/**0.902**	**33.29**/**0.898**	**33.19**/**0.900**

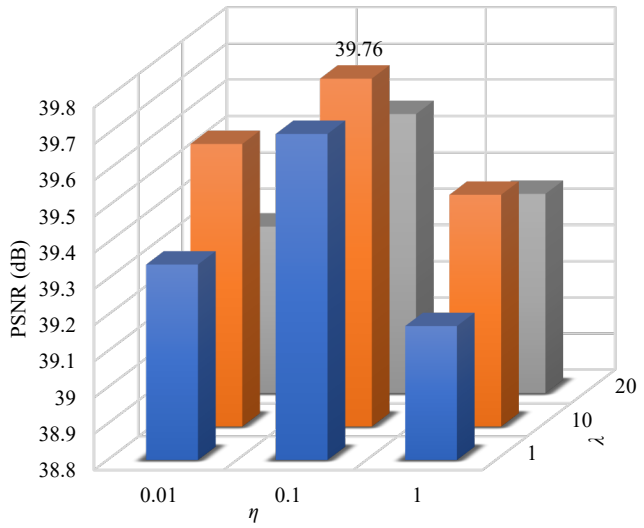

Fig. 6. The mean PSNR values achieved under different combinations of λ and η for 2× stereo SR on Middlebury test set (5 images). The model with $\lambda = 10$ and $\eta = 0.01$ gives the highest mean PSNR value of 39.76 dB.

Therefore, it is necessary to empirically determine the optimal value for the hyper-parameter N that balances the rich feature representation in a certain scale with diversity of feature scales.

To ensure that the receptive field is large enough, the downsampling layers N was empirically set to at least 3. In addition, the spatial resolution of the training samples is not high enough to be downsampled too many times. So, we trained three models with N of 3, 4, and 5, and then tested the performance of these three models with the Middlebury test set (5 images). Table I reports the mean PSNR (dB)/SSIM values for different N values and scale factors. As can be observed, as N increases, the performance of the network also increases first, but then decreases. The network performs best for both left and right views when $N = 4$. Therefore, the number of downsampling layers N is set to 4 by default in the subsequent experiments.

2) Determination of λ and η: In the training phase, since the value of balanced hyper-parameters in the loss function may affect the final performance of the model, we conducted experiments to determine more appropriate hyper-parameter values. We let the value set of λ be $\{1, 10, 20\}$, and the value set of η be $\{0.01, 0.1, 1\}$, generating 9 different combinations of λ and η. We trained 9 models for these different combinations (one for each combination), and tested them on the Middlebury test set for 2× SR. The mean PSNR values achieved under different combinations are shown in Fig. 6. It can be seen from the figure that when the η value is fixed, the model with a λ value of 10 has the best performance, and when the λ is fixed, the model with η set to 0.01 has the best performance. The model trained under $\lambda = 10$ and $\eta = 0.01$ achieved the highest mean PSNR value of 39.76 dB. Therefore, we set $\lambda = 10$ and η=0.01 by default in the subsequent experiments.

D. Ablation Study

1) The Effect of CPAM: The proposed CPAM is used to learn the stereo correspondences and disparity information between the left and right images, so that it facilitates complementary information exchange, and recovers more details. To verify the CPAM's effectiveness, we removed CPAM from the CPASSRnet, retrained the rest of the network with the same training settings, and then tested it with different test sets. As shown in Tab. II, compared with the test results generated by the complete CPASSRnet, if the CPAM is removed, both mean PSNR values and SSIM values of different scale factors on different test sets decline for both views, especially the PSNR values whose maximum drop is 0.16dB. This demonstrates that the proposed CPAM indeed promotes the information interaction between stereo images, and incorporating the additional information can enhance the stereo SR performance. The visual comparison is shown in Fig. 6. Without additional information captured by CPAM, there are two visible artifacts in the corresponding test result. In contrast, with the benefit of CPAM, the result by CPASSRnet is less blurry.

2) The Effect of L1-perceptual Loss: The modified L1-perceptual loss is hypothesized to make the super-resolved images and the corresponding ground truth more perceptually similar, recovering more complete details. The L1-perceptual loss was ablated to verify its effectiveness. We removed the L1-perceptual loss and retrained the CPASSRnet, and the ablation quantitative results on different test sets for $4\times$ stereo SR are also shown in Tab. II. The visual result for $2\times$ stereo SR is shown in Fig. 6. With the L1-perceptual loss removed, the mean PSNR and SSIM values decline, and the SR image is distorated, compared with the integral CPASSRnet. The reason is that the L1-perceptual loss provides an extra training constraint which minimizes the distance between the feature maps (extracted by different types of layers in the pretrained VGG19) of the super-resolved output and that of the ground

TABLE II
MEAN PSNR (DB)/SSIM VALUES ACHIEVED WITH DIFFERENT ABLATIONS BY CPASSRNET ON DIFFERENT TEST SETS FOR 4× STEREO SR.

	without CPAM		without L1-perceptual loss		CPASSRnet	
Dataset	Left	Right	Left	Right	Left	Right
Middlebury (5 images)	33.21/0.897	33.15/0.900	**33.31/0.899**	**33.27/0.901**	**33.37/0.900**	**33.30/0.902**
KITTI 2012 (100 images)	27.40/0.817	27.28/**0.820**	**27.47/0.818**	**27.35/0.820**	**27.50/0.819**	**27.38/0.821**
KITTI 2015 (100 images)	27.68/**0.800**	28.97/0.839	**27.74/0.800**	**29.04/0.840**	**27.78/0.802**	**29.09/0.841**

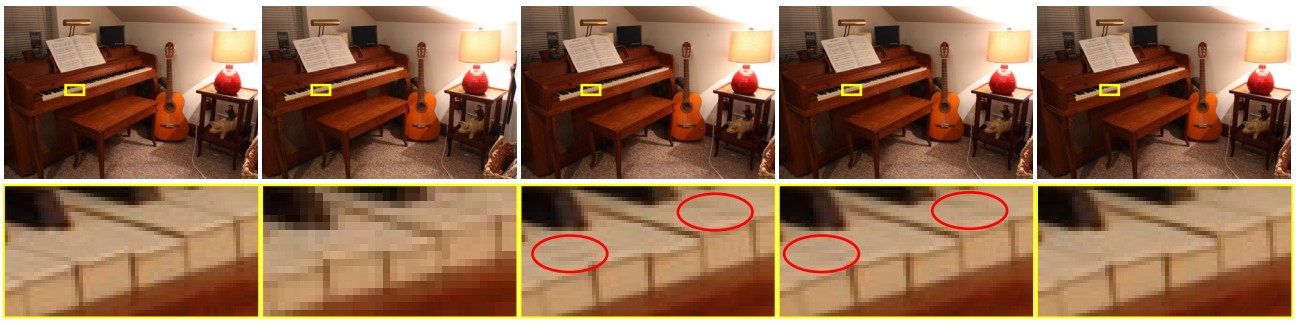

Fig. 7. Visual comparison of different ablating settings for 2× stereo SR on "Piano" of the Middlebury dataset.

truth, which enables the network to recover images more perceptually similar to the ground truth.

3) The Effect of a Multi-scale Training Strategy: A multi-scale training strategy enables a single model to complete super-resolution tasks of multiple scale factors, which is more in line with real-world applications. In this section we report results from ablative experiments to analyze whether multi-scale training can improve the SR performance of the model.

The CPASSRnet models trained separately by scale 2, 3 and 4 are denoted as M_2, M_3 and M_4 respectively, while the model trained by multi-scale (i.e. 2, 3, 4) is denoted as $M_{2,3,4}$. These models are the same except for the training data. We tested these models on the Middlebury test set with integer scales 2, 3, 4 and fractional scales 2.5, 3.5. Figure 8 summarizes the experimental results. The results revealed that the multi-scale training model obtained higher PSNR scores than the single-scale training models on different test scales. (including the fractional scales 2.5 and 3.5 that are not involved in the training step). In particular, in the tests of scales 2, 3 and 4, $M_{2,3,4}$ is at least 2.37 dB better than M_2, M_3 and M_4. This also shows that single-scale models are not good at handling other scales. For example, in the test of scale 2, M_2 gives PSNR of 36.64 dB, but M_3 and M_4 only give 35.77 dB and 33.41 dB, respectively. From these observations, it can be concluded that multi-scale training can effectively boost the SR performance of the model. Intuitively, using ILR images of different scale factors to train a model allows the model to make use of complementary image information across these scales. This can enhance the generalization and expressive ability of the model, so as to better reconstruct high-quality high-resolution images.

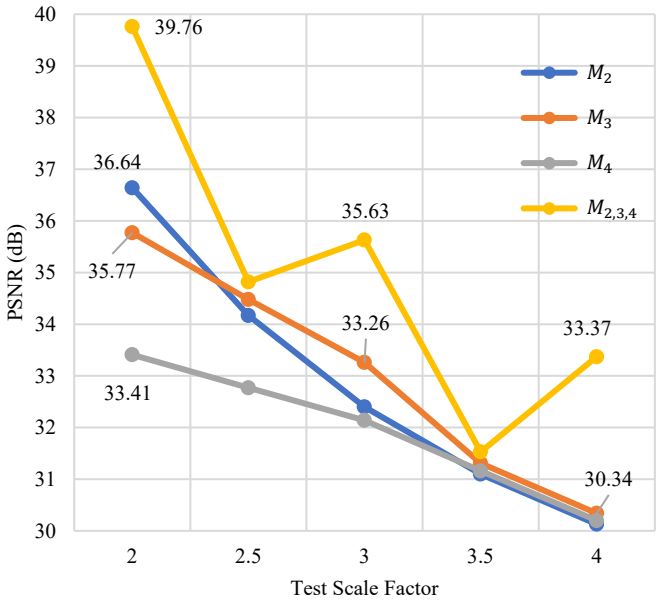

Fig. 8. The mean PSNR values achieved by different models for multiple test scale factors on the Middlebury test set (5 images).

E. Comparison on Images with Horizontal Epipolar Line

In this part, we conduct experiments on CPASSRnet, the conventional bicubic interpolation [55], CNN-based representative SISR methods: SRCNN [22], LapSRN [64], SR-DenseNet [27], and CNN-based advanced methods in stereo SR: StereoSR [8], PASSRNet [9], and then compare their performance based on the test results. All these baselines were trained with original settings illustrated in the original papers and CPASSRnet was trained with default training settings. Since CPASSRnet can super resolve a pair of stereo images

TABLE III
MEAN PSNR (DB)/SSIM VALUES ACHIEVED ON DIFFERENT TEST SETS BY DIFFERENT METHODS FOR DIFFERENT SCALE FACTORS.

		Single Image SR				Stereo Image SR			
Dataset	Scale	Bicubic	SRCNN	LapSRN	SRDenseNet	StereoSR	PASSRNet	Ours (left)	Ours (right)
Middlebury (5 images)	×2	34.77/0.947	38.24/0.971	36.30/0.966	**38.74**/0.969	38.29/0.970	36.30/**0.974**	**39.76**/**0.977**	39.41/0.976
	×3	31.55/0.895	34.17/0.926	33.35/0.924	34.48/0.924	**35.29**/**0.939**	–	**35.63**/**0.938**	35.49/0.939
	×4	29.33/0.836	31.58/0.873	31.52/0.880	31.84/0.873	**32.35**/**0.892**	28.43/0.875	**33.37**/**0.900**	33.30/0.902
KITTI 2012 (100 images)	×2	29.24/0.888	31.16/**0.919**	30.78/0.911	31.32/0.911	31.08/0.912	**31.48**/0.915	**31.55**/**0.920**	31.54/0.922
	×3	26.90/0.825	28.53/0.859	28.51/0.858	28.80/0.852	**28.94**/**0.870**	–	**29.00**/**0.864**	28.89/0.866
	×4	25.32/0.766	26.85/0.805	27.07/**0.812**	27.08/0.801	**27.22**/**0.819**	**27.22**/0.803	**27.50**/**0.819**	27.38/0.821
KITTI 2015 (100 images)	×2	29.41/0.881	**31.08**/**0.916**	30.36/0.898	30.90/0.900	30.77/0.903	28.89/0.893	**31.72**/**0.917**	33.21/0.938
	×3	27.18/0.813	28.54/0.845	28.38/0.835	28.71/0.835	**28.90**/**0.856**	–	**29.27**/**0.852**	30.72/0.886
	×4	25.65/0.752	26.93/0.787	27.04/0.787	27.12/0.781	**27.27**/**0.800**	26.23/0.781	**27.78**/**0.802**	29.09/0.841

simultaneously while the baselines can only handle one view for fair comparison, only the quality and quantitative values of the left view are compared. The quantitative results of the right view obtained by CPASSRnet are separately listed. In addition, the $3\times$ SR results of PASSRNet are not presented because the original paper did not experiment on $3\times$ SR.

1) Qualitative Results: Figure 9 shows the qualitative results for $2\times$ SR and $4\times$ SR. The zoomed-in regions show that SISR methods reconstruct more over-smooth details than the stereo SR methods, which demonstrates the superiority of stereo SR methods. Among the $2\times$ SR results, CPASSRnet can recover sharper edges, making the letters easier to recognize, compared to the other methods. As for $4\times$ SR, the results of these baselines are not only over smooth, but also lack clear details, resulting in the car's wheels appearing incomplete, and also in other places such as detail in the houses.

2) Quantitative Results: The quantitative results achieved on the three previously mentioned test sets are shown in Tab. III. As can be seen, CPASSRnet always gets the highesst PSNR score on all datasets for all scale factors, and gets the highest or second highest SSIM score. For the $2\times$ SR experiment, the test results on the Middlebury dataset show that CPASSRnet is at least 1.02dB higher than other methods in mean PSNR. In terms of SSIM, though CPASSRnet is slightly worse than StereoSR for $3\times$ SR.

It is also worth noting that the right view generated with the left view has a quality comparable to that of the left view, even better than the left view on KITTI 2012 and KITTI 2015 datasets. The reason for this performance is that the proposed CPAM, the L1-perceptual loss and the multi-scale training collaboratively contribute for CPASSRnet to restore higher quality images.

3) Computational Time: In order to compare the speed of previous stereo SR methods with CPASSRnet, we followed StereoSR [8] and resized the Middlebury test set (5 images) to the size of 320×240, and then used these resized images to perform $2\times$ stereo SR tests. The average calculation time of these three methods is: StereoSR (2.964s), PASSRnet (0.185s) and CPASSRnet (0.845s). It can be seen that, in terms of average time for one inference, although CPASSRnet is not the fastest, it can super-resolve both views in one inference. What is more, a single CPASSRnet model can be used for stereo SR task for multiple scale factors. Therefore, considering the convenience and versatility, the time efficiency of CPASSRnet is acceptable.

4) The Ability to Find Stereo Correspondence: The core objective of stereo SR methods is to find the stereo correspondence, to improve the SR performance. To compare the ability of CPASSRnet to find stereo correspondence with other stereo SR methods, we removed the modules which find stereo correspondence (i.e., right view pixel shifting in StereoSR, PAM in PASSRnet, CPAM in CPASSRnet), and used only the left images to retrain them. Apart from removing the modules, all these methods were used the original training settings. The test left results produced by the incomplete models (trained by only left images) are noted as "Left", while the test left results produced by the complete stereo models (trained by left-right pairs) are noted as "Left-Right". As can be seen in Tab. IV, after adding the module which finds stereo correspondence, CPASSRnet achieves the best improvement by 0.12 of PSNR and 0.002 of SSIM. These results indicate that CPASSRnet has a stronger ability to find stereo correspondence, and so can better improve the SR performance.

TABLE IV
COMPARATIVE RESULTS WITH DIFFERENT TYPES OF INPUT FOR $4\times$ STEREO SR ON MIDDLEBURY TEST SET IN TERMS OF MEAN PSNR (DB)/SSIM VALUES.

Input	StereoSR	PASSRnet	Ours
Left	32.28/0.891	28.35/0.874	33.25/0.898
Left-Right	32.35/0.892	28.43/0.875	33.37/0.900
Improvement	0.07/**0.001**	**0.08**/**0.001**	**0.12**/**0.002**

F. Comparison on Images with Irregular Epipolar Line

To verify the superiority of CPASSRnet in dealing with stereo images with irregular epipolar line, 5 pairs of stereo images from ETH3D [7] were randomly selected as a test set. The test models were trained in Sec. IV-E. The qualitative results are shown in Fig. 10. Compared with results using StereoSR and PASSRnet, CPASSRnet reconstructs clean and vivid contours despitethe irregular epipolar line. The quantitative comparative results are shown in Table V: CPASSRnet

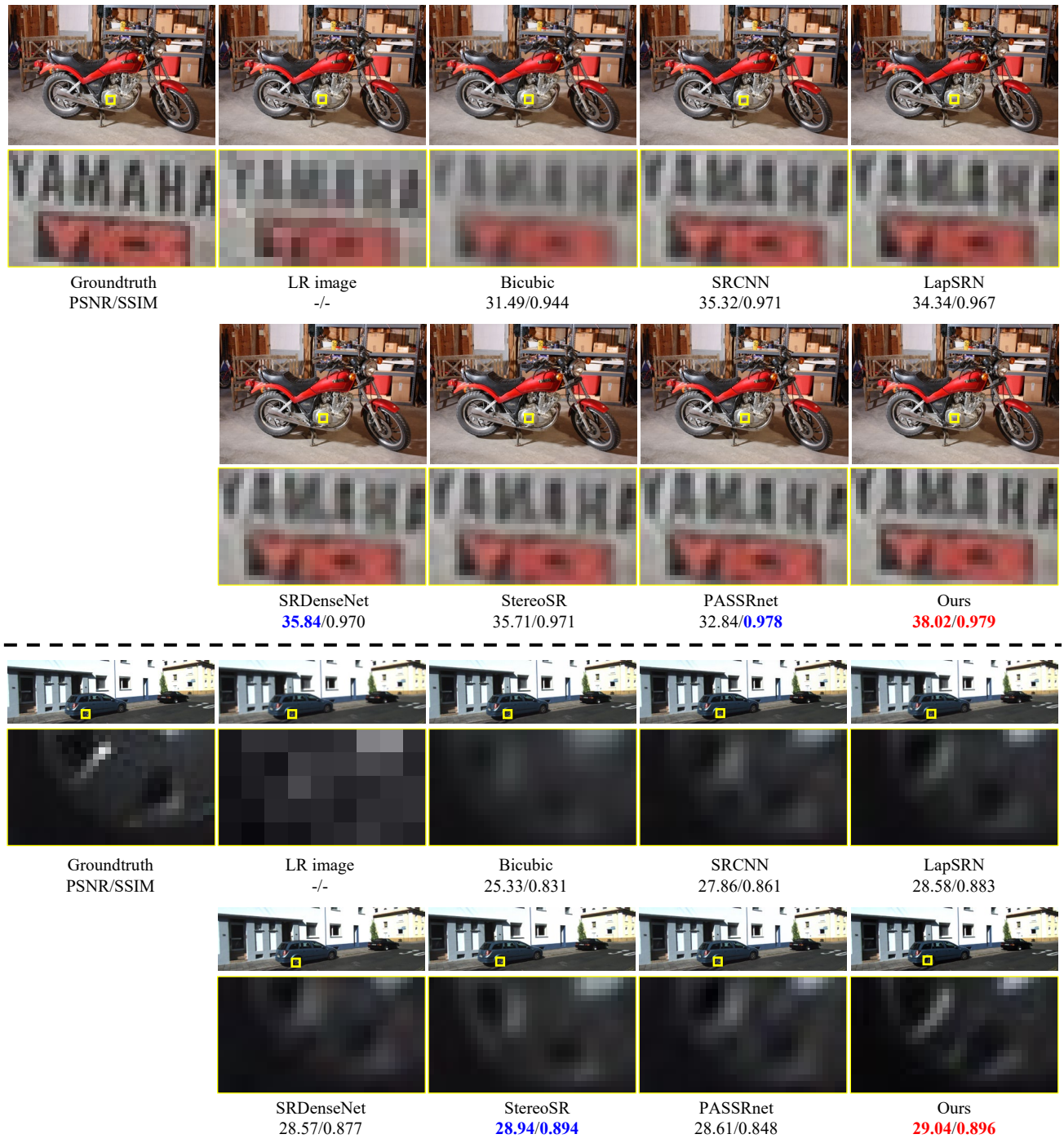

Fig. 9. Visual comparative results for 2× stereo SR (top) and 4× stereo SR (bottom) on horizontal epipolar line datasets.

achieves the best performance for different scale factors, maintaining a performance level consistent with that of stereo images with horizontal epipolar line, or even better. The performance of PASSRnet about mean PSNR/SSIM score is close to being the worst, except for the SSIM value in the 2× stereo SR test. This is because PASSRnet finds stereo correspondence along the horizontal epipolar line, and so it cannot deal with the stereo images with irregular epipolar lines effectively.

G. Noise Test

During shooting, due to many interferences inside and outside the camera, the obtained images are inevitably contaminated by noise. Therefore, the anti-noise ability of the algorithm is also an important indicator to measure the quality of the algorithm. To test the anti-noise ability of CPASSRnet and the stereo SR baselines, Gaussian noise with a mean of 0 and a standard deviation of 3 was added to the LR test images in the Middlebury dataset. We directly input these noisy LR images into the trained models (trained in Sec. IV-E),

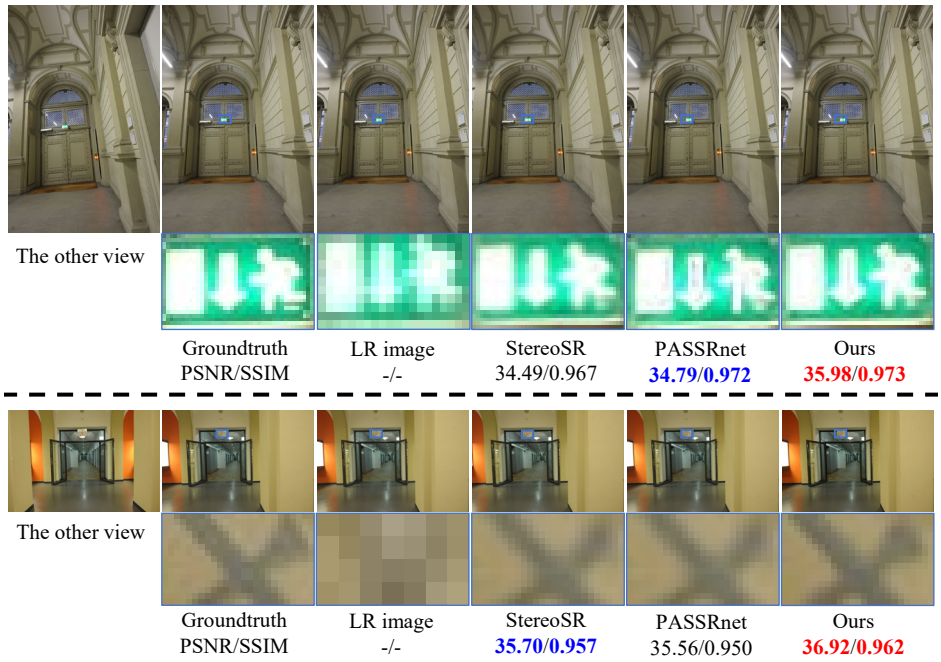

Fig. 10. Visual comparative results for 2× stereo SR (top) and 4× stereo SR (bottom) on irregular epipolar line datasets.

TABLE V
COMPARATIVE RESULTS ACHIEVED ON ETH3D DATASET WITH IRREGULAR EPIPOLAR LINE IN TERMS OF MEAN PSNR (DB)/SSIM VALUES.

Scale	StereoSR	PASSRnet	Ours
×2	32.28/0.891	**37.29/0.975**	**39.22/0.977**
×3	**35.42/0.952**	–	**35.51/0.953**
×4	**33.01/0.925**	30.84/0.911	**33.37/0.931**

and calculate the values of the quantitative indicators for the super-resolved output. The larger the value of these indicators, the stronger the anti-noise ability. The noise test quantitative results are presented in Tab. VI, the proposed CPASSRnet achieves the best performance, exceeding the second-best (i.e., StereoSR) up to 3.22dB in terms of PSNR, and up to 0.051 in terms of SSIM, in 2× stereo SR. The results therefore indicate our method is less sensitive to noise than the other two baselines.

To further investigate the anti-noise stability of these methods, we re-add Gaussian noise with a mean of 0 and a standard deviation of 5 to LR test images in the Middlebury dataset to retest the noise immunity of these models. As a qualitative comparison, Fig. 11 shows a comparison of a small area of representative detail from one of the test images, after processing using CPASSRnet, StereoSR and PASSRnet. It can be seen that CPASSRnet can recover significantly clearer edges and details. The mean PSNR (dB)/SSIM values tested on the noisy LR images are shown in Tab. VII, and CPASSRnet gets the highest mean PSNR/SSIM score again. Also, it is worth noting that the quantitative results of the right view in both of the tests are comparable to the corresponding left view, which demonstrates that CPASSRnet is able to effectively process both views simultaneously. The above tests thus show that our method has stronger and more stable noise immunity. This is because the combination of encoder-decoder structure and max-pooling can effectively filter noise to a significant extent.

TABLE VI
COMPARATIVE RESULTS ACHIEVED ON MIDDLEBURY NOISED TEST SET (THE STANDARD DEVIATION IS 3) IN TERMS OF MEAN PSNR (DB)/SSIM VALUES.

Scale	StereoSR	PASSRnet	Ours (left)	Ours (right)
×2	**32.60/0.851**	25.92/0.799	**35.82/0.902**	32.59/0.902
×3	**31.12/0.820**	–	**33.03/0.864**	33.08/0.865
×4	**28.99**/0.766	25.37/**0.790**	**31.08/0.824**	31.23/0.826

TABLE VII
COMPARATIVE RESULTS ACHIEVED ON MIDDLEBURY NOISED TEST SET (THE STANDARD DEVIATION IS 5) IN TERMS OF MEAN PSNR (DB)/SSIM VALUES.

Scale	StereoSR	PASSRnet	Ours (left)	Ours (right)
×2	**29.99/0.730**	21.08/0.598	**32.56/0.800**	32.54/0.800
×3	**29.05/0.702**	–	**30.68/0.765**	30.74/0.767
×4	**27.45**/0.651	23.58/**0.690**	**29.14/0.724**	29.09/0.725

V. CONCLUSION

Stereo super-resolution techniques have many application scenarios. For example, to dynamically adjust the resolution of stereo images, so that they fit the hardware conditions of a particular 3D display device, promoting a better display effect. Therefore, it is important to improve the performance of stereo super-resolution techniques. In this paper, we have proposed a cross parallax attention stereo super-resolution network (CPASSRnet) to perform stereo SR at multiple scale

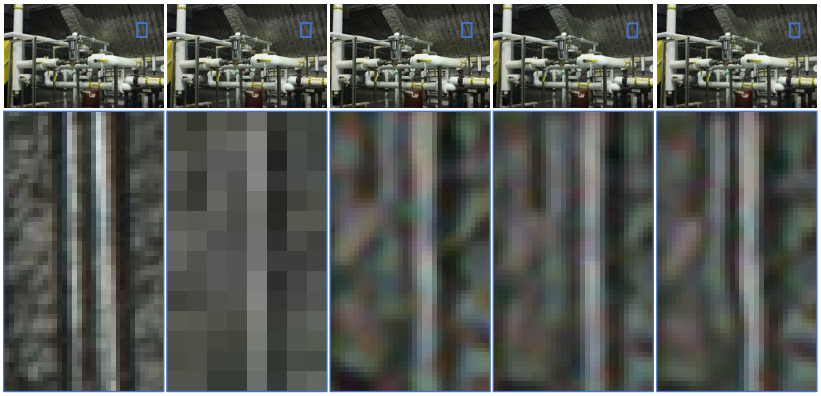

Fig. 11. Comparative results for 4× SR on noised "Pipes" of the Middlebury [60] dataset.

factors, for both left and right views, with a single model. Using our proposed cross parallax attention module (CPAM), which is able to capture the global stereo correspondence of additional information for each view relative to another view, CPASSRnet can handle stereo images with large disparity and different types of epipolar lines, unlike existing methods.

Firstly, we conducted a sensitivity analysis to determine the values of hyper-parameters. Then the ablation experiments were carried out to illustrate the superiority of the proposed method, which benefits from the cross parallax attention module, the L1-perceptual loss and the multi-scale training strategy. We also conducted a series of experimental comparisons with state-of-the-art methods. Compared with the state-of-the-art stereo SR methods, on test sets with horizontal epipolar lines, results show that our method performs best in terms of the PSNR index, improving on the next best by at least 0.06. On the SSIM index, most of the test results showed that our method is comparable with the best of the other methods, lagging at most by 0.006. In the experiment of verifying the ability to find stereo correspondence, our CPAM showed stronger performance, and can improve the SR performance by mean PSNR/SSIM of 0.12/0.002. On test sets with irregular epipolar lines, our method outperforms the baselines, which indicates that our method is more suitable for processing the stereo images with irregular epipolar lines. Finally, in order to test the robustness of the proposed method, noise tests were conducted. The experiments illustrated that after adding Gaussian white noise with a standard deviation of 3 or 5, our method is still stable and can achieve the best performance, exceeding the second best by at least 1.63/0.034 in terms of mean PSNR/SSIM, which verifies that CPASSRnet is less sensitive to noise compared to the stereo SR baselines. Thus, over a range of metrics, our method outperforms state-of-the-art techniques, and represents a promising improvement for stereo SR applications.

Although our method can achieve better performance and can perform stereo SR task for multiple scaling factors in one model, it is also a more complex model with a relatively large number of parameters. However, we firmly believe that there are better ways to integrate additional information other than CPAM. We will study how to achieve comparable or even better performance while simplifying model complexity in the future, and explore better ways to integrate additional information.

REFERENCES

[1] S. Zhou, J. Zhang, W. Zuo, H. Xie, J. Pan, and J. S. Ren, "Davanet: Stereo deblurring with view aggregation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 996–11 005.

[2] P. Li, X. Chen, and S. Shen, "Stereo r-cnn based 3d object detection for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7644–7652.

[3] J. Li, L. Wei, F. Zhang, T. Yang, and Z. Lu, "Joint deep and depth for object-level segmentation and stereo tracking in crowds," *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2531–2544, 2019.

[4] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, "Stereoscopic neural style transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6654–6663.

[5] H. Cheng, J. Zhang, Q. Wu, and P. An, "A computational model for stereoscopic visual saliency prediction," *IEEE Transactions on Multimedia*, vol. 21, no. 3, pp. 678–689, 2018.

[6] D. Scharstein and C. Pal, "Learning conditional random fields for stereo," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[7] T. Schops, J. L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3260–3269.

[8] D. S. Jeon, S.-H. Baek, I. Choi, and M. H. Kim, "Enhancing the spatial resolution of stereo images using a parallax prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1721–1730.

[9] L. Wang, Y. Wang, Z. Liang, Z. Lin, J. Yang, W. An, and Y. Guo, "Learning parallax attention for stereo image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 250–12 259.

[10] X. Ying, Y. Wang, L. Wang, W. Sheng, W. An, and Y. Guo, "A stereo attention module for stereo image super-resolution," *IEEE Signal Processing Letters*, vol. 27, pp. 496–500, 2020.

[11] J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[12] G. Freedman and R. Fattal, "Image and video upscaling from local self-examples," *ACM Transactions on Graphics (TOG)*, vol. 30, no. 2, pp. 1–11, 2011.

[13] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 349–356.

[14] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Computer graphics and Applications*, vol. 22, no. 2, pp. 56–65, 2002.

[15] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 1. IEEE, 2004, pp. I–I.

[16] M. Yang and Y. F. Wang, "A self-learning approach to single image super-resolution," *IEEE Transactions on Multimedia*, vol. 15, no. 3, pp. 498–508, 2013.

[17] Z. Zhu, F. Guo, H. Yu, and C. Chen, "Fast single image super-resolution via self-example learning and sparse representation," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2178–2190, 2014.

[18] Z. Xiong, X. Sun, and F. Wu, "Robust web image/video super-resolution," *IEEE transactions on image processing*, vol. 19, no. 8, pp. 2017–2028, 2010.

[19] K. I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 6, pp. 1127–1133, 2010.

[20] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[21] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *2008 IEEE conference on computer vision and pattern recognition*. IEEE, 2008, pp. 1–8.

[22] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[23] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[25] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.

[26] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646–1654.

[27] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4799–4807.

[28] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144.

[29] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2472–2481.

[30] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 252–268.

[31] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 517–532.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[33] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 286–301.

[34] J. Liu, W. Zhang, Y. Tang, J. Tang, and G. Wu, "Residual feature aggregation network for image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2359–2368.

[35] C. Tian, Y. Xu, W. Zuo, B. Zhang, L. Fei, and C.-W. Lin, "Coarse-to-fine cnn for image super-resolution," *IEEE Transactions on Multimedia*, 2020.

[36] J. Guo, S. Ma, J. Zhang, Q. Zhou, and S. Guo, "Dual-view attention networks for single image super-resolution," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2728–2736.

[37] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[38] Y. Mei, Y. Fan, Y. Zhou, L. Huang, T. S. Huang, and H. Shi, "Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5690–5699.

[39] B. Niu, W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, and H. Shen, "Single image super-resolution via a holistic attention network," in *European Conference on Computer Vision*. Springer, 2020, pp. 191–207.

[40] A. V. Bhavsar and A. Rajagopalan, "Resolution enhancement for binocular stereo," in *2008 19th International Conference on Pattern Recognition*. IEEE, 2008, pp. 1–4.

[41] H. Park, K. M. Lee, and S. U. Lee, "Combining multi-view stereo and super resolution in a unified framework," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2012, pp. 1–4.

[42] W. Song, S. Choi, S. Jeong, and K. Sohn, "Stereoscopic image super-resolution with stereo consistent feature." in *AAAI*, 2020, pp. 12 031–12 038.

[43] B. Yan, C. Ma, B. Bare, W. Tan, and S. C. Hoi, "Disparity-aware domain adaptation in stereo image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 179–13 187.

[44] A. P. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A decomposable attention model for natural language inference," *arXiv preprint arXiv:1606.01933*, 2016.

[45] J. Cheng, L. Dong, and M. Lapata, "Long short-term memory-networks for machine reading," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 551–561.

[46] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International Conference on Machine Learning*, 2019, pp. 7354–7363.

[47] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[48] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 510–519.

[49] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[50] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 11 065–11 074.

[51] Y. Qiao, Y. Liu, X. Yang, D. Zhou, M. Xu, Q. Zhang, and X. Wei, "Attention-guided hierarchical structure aggregation for image matting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 676–13 685.

[52] M. Yin, Z. Yao, Y. Cao, X. Li, Z. Zhang, S. Lin, and H. Hu, "Disentangled non-local neural networks," in *European Conference on Computer Vision*. Springer, 2020, pp. 191–207.

[53] J. Li, X. Liu, W. Zhang, M. Zhang, J. Song, and N. Sebe, "Spatio-temporal attention networks for action recognition and detection," *IEEE Transactions on Multimedia*, 2020.

[54] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, and Q. Dai, "Stat: spatial-temporal attention mechanism for video captioning," *IEEE transactions on multimedia*, vol. 22, no. 1, pp. 229–241, 2019.

[55] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE transactions on acoustics, speech, and signal processing*, vol. 29, no. 6, pp. 1153–1160, 1981.

[56] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.

[57] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[58] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[59] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, vol. 1. IEEE, 2003, pp. I–I.
[60] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *German conference on pattern recognition*. Springer, 2014, pp. 31–42.
[61] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving," in *Proc. CVPR*, pp. 3354–3361.
[62] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3061–3070.
[63] D. P. K. JLB, "Adam: A method for stochastic optimization," in *3rd international conference for learning representations, San Diego*, 2015.
[64] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 624–632.

Canqiang Chen received the B.Eng. degree from Guang-dong University of Technology, Guangzhou, China, in 2019. He is currently pursuing the M.Eng. degree with the School of Electronic and Information Engineering, South China University of Technology (SCUT), Guangzhou, China. His research interests are in computer vision, machine learning, and image processing.

Chunmei Qing received a bachelor of science degree in 2003 and received a doctorate degree in Information and Computing Science from Sun Yat-sen University. She received her Ph.D. degree in Electronic Imaging and Media Communication from the University of Bradford, UK, in 2009. Then, she worked as a postdoctoral researcher at University of Lincoln in the United Kingdom. Since 2013, she has been an associate professor in the School of Electronics and Information Engineering, South China University of Technology (SCUT), China. Her main research interests include image/video processing, computer vision, pattern recognition and machine learning.

Xiangmin Xu received a doctorate degree from South China University of Technology (SCUT), China. He is currently a full-time professor in the School of Electronics and Information Engineering, SCUT. His current research interests include image/video processing, human-computer interaction, computer vision and machine learning.

Patrick Dickinson has studied at the University of Southampton (BSc Physics), University of Oxford (MSc Computation), and completed his PhD in Computer Vision at the University of Lincoln in 2008. He is currently an Associate Professor and the acting director of the intLab research group in the School of Computer Science at the University of Lincoln, UK. His main research interests include human computer interaction, and user experiences in virtual reality and games.